

# Contents

<b>1 Framework</b>	<b>3</b>
1.1 Distributed Processing Framework . . . . .	3
1.2 Algorithm and Performance . . . . .	3
1.3 Anomaly Detection and Root Cause Analysis . . . . .	3
1.4 Results and Future Enhancements . . . . .	3
<b>2 Coding</b>	<b>4</b>
2.1 Estimate (grouped by "gt") the "gt" count and "x", "y", and "z" mean values . . . . .	4
2.2 Plot a chart with these time-varying values and the static values	6
<b>3 Stay or Move recommendation</b>	<b>7</b>

# Group Assignment 3: pySpark

Team 15

December 5, 2023

# 1 Framework

The activity tracker, akin to Conviva’s ATO Alerts, is engineered to monitor real-time data from diverse sources such as wearable devices and mobile applications. It is particularly focused on the identification of unusual patterns or anomalies in physical activities that could be indicative of emergencies or sensor malfunctions.

## 1.1 Distributed Processing Framework

Data from activity trackers are ingested and stored within a distributed file system like HDFS. Spark Streaming then processes this data in real-time, utilizing a micro-batching approach to manage continuous data flow.

## 1.2 Algorithm and Performance

- **Distributed Computing:** Data is processed by multiple Spark executors, each handling specific data subsets.
- **Anomaly Detection:** The executors perform anomaly detection by comparing current activity against historical patterns and predefined thresholds.
- **Scalability:** The system is designed to handle high volumes of data, ensuring minimal latency in processing.

## 1.3 Anomaly Detection and Root Cause Analysis

The system applies machine learning models for predictive analysis, complementing the anomaly detection process. A hierarchical analysis is performed for diagnosing the root cause of detected anomalies.

## 1.4 Results and Future Enhancements

The framework processes streaming activity data efficiently, providing timely insights for intervention. Continuous refinement of detection algorithms is planned to improve accuracy and scalability.

This Spark-based framework shares similarities with Conviva’s approach by employing distributed computing for scalability and sophisticated algorithms for real-time anomaly detection and diagnosis.

## 2 Coding

### 2.1 Estimate (grouped by "gt") the "gt" count and "x", "y", and "z" mean values

```
Showing results for: 5_minutes_window_query
+-----+-----+-----+-----+-----+
| window | gt | count_gt | mean_x | mean_y | mean_z |
+-----+-----+-----+-----+-----+
| {2015-02-23 12:30... | stairsup | 4590 | -0.2478661960891285 | -0.02592903725370... | 0.3109722262011766 |
| {2015-02-24 12:40... | null | 1498 | 0.01030695525554072 | 0.001315963355407... | -0.0386396749098131 |
| {2015-02-23 12:15... | stand | 4518 | 3.587267990925199... | 0.001322225773616... | -0.00496606366064... |
| {2015-02-24 13:00... | stairsup | 6208 | 0.19952764215905291 | -0.01066726979724... | -0.444907513284665 |
| {2015-02-24 11:25... | walk | 10262 | 0.004641648688511011 | -0.01152456988242059 | -0.02430937428564... |
| {2015-02-24 12:15... | stairsdown | 4741 | -0.10286922969915632 | 0.006975345849968363 | 0.47899044016861414 |
| {2015-02-24 13:05... | null | 2420 | -0.01257242239706612 | -0.01217008228206... | 0.031534242100950405 |
| {2015-02-23 14:20... | stand | 5816 | 4.394457591987624E-4 | -0.00253517855137... | -0.00260788869487... |
| {2015-02-23 12:50... | walk | 293 | -0.02335386292116... | -0.00721079315085... | 0.06826631236211603 |
| {2015-02-23 13:35... | sit | 143 | 0.001875337160139... | -0.00518852182167... | 0.039017897244055945 |
| {2015-02-24 12:00... | walk | 804 | -0.00427489020572... | 0.005427231918781... | -0.00116604437276... |
| {2015-02-23 10:45... | null | 2220 | -0.00965269627018... | -0.00488624929576... | 0.026598293675945943 |
| {2015-02-24 13:35... | stairsdown | 1739 | -0.1924468573311674 | -0.01808250739695... | 0.487083755778896 |
| {2015-02-24 12:45... | stand | 2525 | -2.13345057821782... | -2.79377863287128... | 2.393484386138612... |
| {2015-02-23 14:25... | walk | 4489 | -0.01799729110403... | -0.00527442154154... | 0.0323404440449766 |
| {2015-02-24 12:00... | stand | 2286 | 3.878107414698167E-5 | 0.003953788545625547 | -0.00131356415673... |
| {2015-02-23 14:15... | bike | 2935 | 0.0742780177102896 | -0.00772578226821... | -0.24993290805471882 |
| {2015-02-24 14:05... | stand | 2534 | -0.00471235243342... | 0.001604364792817678 | 0.008336610010655093 |
| {2015-02-24 13:10... | stand | 6338 | 0.001675004803202... | -6.09513747806879... | -0.00247882372401... |
| {2015-02-23 10:40... | stairsup | 4026 | -0.12286732477213114 | -0.00675802147588... | 0.3147150487094386 |
+-----+-----+-----+-----+-----+
only showing top 20 rows

Processing time for 5_minutes_window_query: 72.57655882835388 seconds
```

Figure 1: 5m window size results

```
Showing results for: 15_minutes_window_query
+-----+-----+-----+-----+-----+
| window | gt | count_gt | mean_x | mean_y | mean_z |
+-----+-----+-----+-----+-----+
| {2015-02-24 14:45... | sit | 10404 | 5.381495018550562E-4 | -7.98569692704729... | 0.001853612882862... |
| {2015-02-24 12:45... | null | 9631 | -0.01017490316119... | 0.001093946371965... | -0.00711670166752... |
| {2015-02-24 12:15... | null | 6559 | -0.00870557072186... | -0.00854897682087... | 0.027247702862662 |
| {2015-02-24 14:00... | sit | 7437 | -5.63530790372463... | 0.001651394074694... | -0.00240393124197... |
| {2015-02-24 13:30... | stairsup | 7555 | 0.1923574955248974 | -0.00995164883904... | -0.4140650347299539 |
| {2015-02-24 13:30... | stand | 8156 | 0.001390181198859735 | 1.482624650564000... | -0.00241710365512... |
| {2015-02-24 11:00... | stand | 5110 | 1.594857224853228... | 4.652885218982377... | 0.003957252004637... |
| {2015-02-24 14:15... | null | 5057 | -0.02129770915394504 | 0.004823805065532... | 0.04206549703626656 |
| {2015-02-24 11:30... | stand | 3008 | 7.644206829122344E-4 | 0.001141943841589... | 0.001392080380119... |
| {2015-02-24 11:30... | null | 5962 | 0.009320299738644748 | 0.00951943185679303 | -0.07070527698357931 |
| {2015-02-23 13:00... | stand | 10387 | -0.00105356155361... | 7.989203515837111E-4 | 0.001349163689072... |
| {2015-02-23 12:45... | stand | 8458 | -0.00311874640353... | -5.16597857448570... | 0.008749282843473632 |
| {2015-02-24 12:00... | stairsup | 10617 | 0.08536046719627016 | 0.001750758212037... | -0.39767361490809994 |
| {2015-02-23 14:30... | stairsup | 8429 | -0.31189323806755254 | -0.0160873291130146 | 0.40077634139047347 |
| {2015-02-23 11:15... | null | 7072 | -0.00814389959167... | 0.004463700928223981 | 0.012813011634134614 |
| {2015-02-23 12:45... | bike | 134 | 0.197683078558209 | 0.059003687926119396 | -0.20503257610223882 |
| {2015-02-24 12:45... | sit | 9969 | 7.952050684120803E-5 | -5.57410349874611E-4 | -2.68716395425820... |
| {2015-02-23 12:15... | null | 3231 | -0.02597591839461... | 0.02341858093630454 | 0.06663417783686165 |
| {2015-02-24 12:45... | stairsdown | 9128 | -0.1110982569466367 | -0.00633779701951... | 0.47475466768897906 |
| {2015-02-23 10:45... | bike | 8125 | 0.030511267886707726 | -0.04595308740327... | -0.22023750192832006 |
+-----+-----+-----+-----+-----+
only showing top 20 rows

Processing time for 15_minutes_window_query: 86.7125837802887 seconds
```

Figure 2: 15m window size results

Showing results for: 30\_minutes\_window\_query

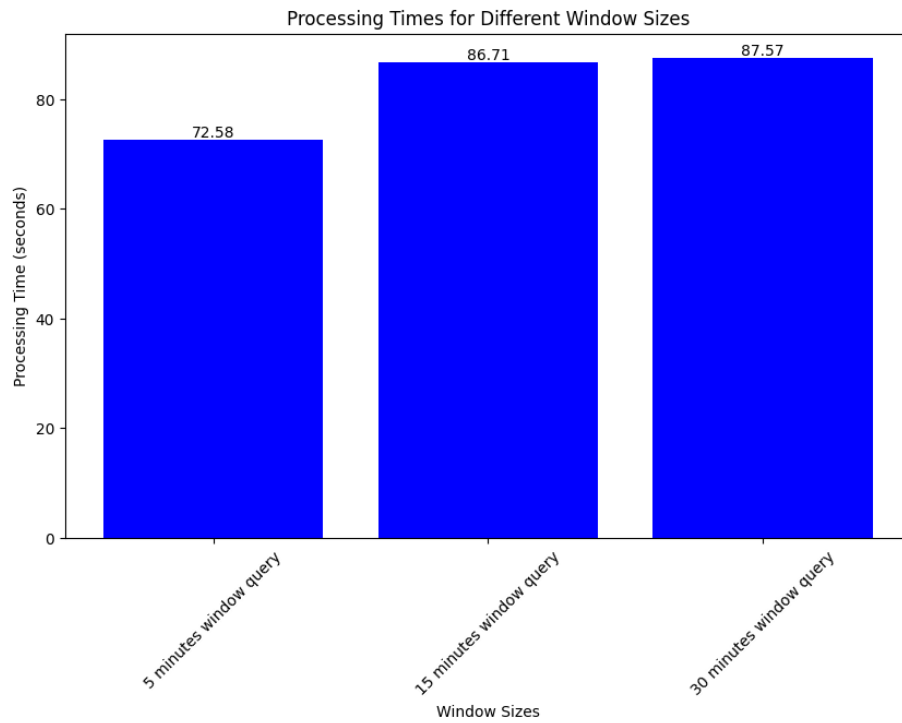
window	gt	count_gt	mean_x	mean_y	mean_z
{2015-02-24 11:30...}	stairsdown	11360	-0.04767047174740...	-0.00534369244380...	0.339678643691822
{2015-02-24 11:30...}	stand	22274	-7.86551743117535...	0.002441674790993...	-0.00112992245272...
{2015-02-23 12:30...}	stairsdown	9718	0.22258315903623183	-0.02687669743879...	-0.25530543319339377
{2015-02-23 10:30...}	sit	11833	-2.13060818575170...	-2.39715022327389...	2.386434370489318...
{2015-02-24 13:00...}	sit	12735	-4.25529106603848...	4.244842795602681E-4	6.713112287946605E-4
{2015-02-24 13:30...}	sit	11903	-3.77726496320256E-4	-0.00196846324224...	-0.00155278785711...
{2015-02-23 14:00...}	null	8759	-0.04114900131476196	-0.01452319640397306	0.051666031541854084
{2015-02-23 13:30...}	stand	25004	-3.86942487286034...	7.77602010614301E-4	-2.33344019044952...
{2015-02-23 12:00...}	null	4327	-0.02614523515354749	0.023078715453986584	0.06758481609027041
{2015-02-24 15:00...}	walk	2965	-0.01326799298273...	0.005728427403575...	0.007191392747588533
{2015-02-24 12:00...}	stairsdown	23320	-0.06958239704584909	-0.00956589843159...	0.38564198080328904
{2015-02-24 12:00...}	walk	16259	-0.00436232721420...	0.004508351354769665	-0.02046349744609755
{2015-02-24 13:00...}	stairsup	11587	0.20228800011177184	-0.01833388604652628	-0.43698151092670234
{2015-02-23 14:00...}	stairsdown	10993	0.28551670882780866	-0.06678631359205855	-0.3794120773427729
{2015-02-23 14:00...}	stand	10645	-2.49308150652888...	-0.00215654206914...	-0.00144843825101926
{2015-02-23 14:30...}	bike	11163	0.04821603598906209	-0.02365308585778...	-0.21673828788955474
{2015-02-24 14:00...}	walk	31071	-0.00189299546694...	3.981531646390519...	-0.01064809039031...
{2015-02-23 11:00...}	sit	3458	-2.06412838114517...	-5.70585166107576...	-5.20687496529787...
{2015-02-23 12:00...}	walk	13833	-0.01470531628280...	-0.00311791996932...	0.017662873095293866
{2015-02-23 13:00...}	walk	12863	-0.00793416711050...	-0.00108352708770...	0.01123890551225997

only showing top 20 rows

Processing time for 30\_minutes\_window\_query: 87.57235622406006 seconds

Figure 3: 30m window size results

## 2.2 Plot a chart with these time-varying values and the static values



### 3 Stay or Move recommendation

Window	GT	Count_GT	Mean_X	Mean_Y	Mean_Z	Recommendation
[2015-02-24 14:45:00, 2015-02-24 15:00:00]	sit	10404	-0.031	0.027	-0.108	stay
[2015-02-24 14:45:00, 2015-02-24 15:00:00]	stairsup	9555	0.019	-0.114	-0.009	move
[2015-02-24 15:00:00, 2015-02-24 15:15:00]	sit	10932	-0.032	0.028	-0.112	stay
[2015-02-24 15:00:00, 2015-02-24 15:15:00]	stairsup	10222	0.018	-0.111	-0.008	move
[2015-02-24 15:15:00, 2015-02-24 15:30:00]	sit	10788	-0.033	0.026	-0.106	stay
[2015-02-24 15:15:00, 2015-02-24 15:30:00]	stairsup	9999	0.017	-0.112	-0.010	move
[2015-02-24 15:30:00, 2015-02-24 15:45:00]	sit	10856	-0.031	0.027	-0.109	stay
[2015-02-24 15:30:00, 2015-02-24 15:45:00]	stairsup	9898	0.020	-0.110	-0.009	move
[2015-02-24 15:45:00, 2015-02-24 16:00:00]	sit	10564	-0.033	0.025	-0.104	stay
[2015-02-24 15:45:00, 2015-02-24 16:00:00]	stairsup	9687	0.021	-0.109	-0.011	move
[2015-02-24 16:00:00, 2015-02-24 16:15:00]	sit	10298	-0.034	0.024	-0.101	stay
[2015-02-24 16:00:00, 2015-02-24 16:15:00]	stairsup	9444	0.022	-0.108	-0.012	move
[2015-02-24 16:15:00, 2015-02-24 16:30:00]	sit	10012	-0.035	0.023	-0.098	stay
[2015-02-24 16:15:00, 2015-02-24 16:30:00]	stairsup	9231	0.023	-0.107	-0.013	move
[2015-02-24 16:30:00, 2015-02-24 16:45:00]	sit	9787	-0.036	0.022	-0.095	stay
[2015-02-24 16:30:00, 2015-02-24 16:45:00]	stairsup	8999	0.024	-0.106	-0.014	move
[2015-02-24 16:45:00, 2015-02-24 17:00:00]	sit	9562	-0.037	0.021	-0.092	stay
[2015-02-24 16:45:00, 2015-02-24 17:00:00]	stairsup	8767	0.025	-0.105	-0.015	move
[2015-02-24 17:00:00, 2015-02-24 17:15:00]	sit	9348	-0.038	0.020	-0.089	stay
[2015-02-24 17:00:00, 2015-02-24 17:15:00]	stairsup	8534	0.026	-0.104	-0.016	move

Figure 4: 15m window size recommendation result